CNA

A REFERENCE-CONNECTING TECHNIQUE FOR AUTOMATIC INFORMATION CLASSIFICATION AND RETRIEVAL

By M.J. Greene

OEG Research Contribution No. 77

CONTRACT N00014-67-C-0211

Research Contribution

OPERATIONS EVALUATION GROUP
Center for Naval Analyses
THE FRANKLIN INSTITUTE
WASHINGTON, D.C. 20350

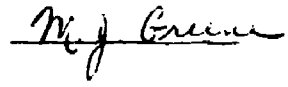
OEG RESEARCH CONTRIBUTION NO. 77

Operations Evaluation Group

CENTER FOR NAVAL ANALYSES

A REFERENCE CONNECTING TECHNIQUE FOR AUTOMATIC INFORMATION CLASSIFICATION AND RETRIEVAL

By M.J. Greene



10 March 1967

Work conducted under contract N00014-67-C-0211

ABSTRACT

A recent study of command information flow associated with the Dominican Republic coup of April-May 1965 introduced an analytical tool for identifying deficiencies in the flow and use of information which appears to have considerable potential as a general technique for information retrieval. Naval messages are associated with each other through their formal references in a manner analogous to the concept of "joining" as defined in a recent publication in directed graph theory by Harary, Norman, and Cartwright (Structural Models). "Reference-connected sets" are then constructed from message traffic dealing with the coup and are found to uniquely identify operational events during the crisis. Such sets (and subsets if further refinement is desired) can be obtained in real-time and a very simple method is demonstrated which automatically classifies messages as they enter the system. This technique, if applied to a library system, avoids both the problem of describing the subject covered in a document and the problem of integrating new subject matter into a predetermined classification code. It is a user-oriented system which combines static coupling methods used in the Technical Information Project at MIT and converts them into a dynamic process for both classification and retrieval. Thus, it concentrates on the evolution of the subject as represented by referenced documents upon which any "new" information is based and it has the additional advantage of reflecting the information flow among scientists in any field.

(REVERSE BLANK)

INTRODUCTION

Computer technology is purported to have caused an "information explosion" which has resulted in the evolution of a new science. This new "information system science,"[1] which is a yet undefined composite of many established divisions of knowledge such as engineering, mathematics, logic, linguistics, operations research, management science, library science, etc., has as one of its subdivisions the area of "information retrieval." So while scientists in the more established disciplines continue to search for solutions to the "information retrieval problem" which will suit their own specific needs, basic research is beginning to be conducted which is not directed toward any particular user or system.

Military "command and control" is another specialty which poses its own problems of information retrieval, not only for those involved in the design of command support systems but for those who attempt to develop models to determine functional requirements for such systems. For the past few years, the Operations Evaluation Group of the Center for Naval Analyses has performed several information flow analyses as a contribution to the latter objective which have been aimed at finding out "who talks to whom, about what, and how effectively" in a wide range of operational situations featuring the involvement of naval forces and commands. In a current analysis of the Dominican Republic crisis of April - May 1965, an analytical tool was developed for describing information flow throughout the naval chain-of-command which also appears to have considerable potential as a technique for information retrieval, not only in operational control centers, but for any system which has as one of its functions the selection of information from a store. This technique is presented here as a contribution to the general study and with the hope that its potential utility will be evaluated by other scientists in other fields.

DISCUSSION

Message traffic among higher-echelon commands during the early part of a crisis situation is extremely difficult to classify. This is because such communications do not generally fall into categories which deal with specific predetermined military tasks, but are much less precisely defined, less routine, and consist primarily of the exchanges of information along with recommendations, advice, etc., which are necessary before any tactical systems can be put into effect. By the same token, these communications are difficult to retrieve in any formatted sense because the unexpected, evolving, and interdependent nature of the information places an even greater emphasis upon natural language communication.

[1] Footnotes are listed on page 11.

In an attempt to avoid the inadequacies inherent in any classification system while at the same time recognizing the fact that as the amount of available information grew there was a parallel need for a more precise way to retrieve specific data[2], a technique was developed for associating messages with each other which required no interpretation of the subject content of the messages. This technique is based upon the thesis that if a message referenced a previous message, the previous message must have influenced that message in some way. For example, a message might say, "This is in answer to your question in reference A..." Often a message referenced a previous message which referenced a yet earlier message. Still other connections of messages through their references are possible. In figure 1, if each circled number represents a message and if an arrow from, say ② to ① means ② referenced ①, then we can interpret the figure as follows: message ② references message ①, message ④ references message ② but also references message ③, and message ⑤ is another message which references message ③. Thus we can speak of a "reference-connected" set S = {①, ②, ③, ④, ⑤} of messages - that is, a set of messages which are connected in any way through their references.[3]

FIG. 1: A REFERENCE-CONNECTED SET OF MESSAGES

This concept of reference-connecting is analogous to the one of "joining" as defined in a recent publication in directed graph theory by Harary, Norman, and Cartwright (Structural Models) and more explicit (although not the most fundamental) definitions can be given:

Definition 1. A directed line from message ① to message ② means that ① referenced ②. (①→②)

Definition 2. A path from message ① to message ② exists if ② can be reached from ① through a sequence of directed lines. (①→③→④→②) [This amounts basically to the tracing of references.]

Definition 3. A semipath from message ① to message ② exists if ② can be joined to ① through a sequence of lines in which the direction is ignored. (①→③←④→②) [Note that every path is a semipath but not every semipath is a path.]

Definition 4. A reference-connected set of messages is a set of messages which are joined through semipaths.

Definition 5. A subset S' of a reference-connected set S is a set whose messages are messages in S.

When messages dealing with the Dominican Republic crisis were put in reference-connected sets, these sets in most cases uniquely identified particular events during the crisis. One set which was constructed from crisis-related message traffic found in files at three command headquarters contained 105 distinct messages which dealt with the preparations for landing airborne troops. Other sets of messages represented the communications related to other events such as the provision of medical supplies, the preparation of evacuation lists, the sending of surgical teams, and others. All of these events were represented by unique message sets in the investigated files of crisis-related traffic.

Reference-connected sets proved to be valuable tools in analyses of command information flow and for analyses of the operations they describe. Deficiencies in flows and use of information are much more easily identified when focus is placed upon a specific event represented by communications throughout an entire command structure. Such sets can also be displayed in a manner which shows what information regarding an operational event was available to various commands, and what lapses there were in the message distribution logic. Figure 2 illustrates an actual reference-connected set which corresponds to the event of sending food to the Dominican Republic.[4] Nineteen commands were involved as originators or addressees in the 15 reference-connected messages found at one command headquarters, although only six commands were directly involved in the operation. These six commands are labeled A, B, C, D, E, and F, and are represented on a vertical axis. The messages are numbered on the horizontal axis and are placed on the axis (representing increasing time in the positive direction) according to the time they were originated. Horizontal arrows connecting messages indicate that Message ⓐ was referenced by (←—) Message ⓑ . Vertical arrows between commands indicate that Command X addressed (—→) Command Y for action, and a bar at Command Z (—) indicates that Z was an information addressee on this message.

Figure 2 - A Narrative. The first message was sent from Command A to a subordinate Command B for action and to Commands C and D for information. It requested information regarding the logistics of unloading food in the Dominican Republic. In the next message, Command B referenced message ① and requested advice from Administrative Commands D and E. This message was not addressed to Command C who sent message ③ to his superior Command B in which he stated that he had no knowledge of the subject except as contained in the reference. Command A, who was still waiting for information, repeated his request to Command B in message ④ . The next two messages which were sent to Command C both directed that he "take reference for action ASAP (as soon as possible)." Command E referenced message ② (which was not originally sent to Command C) and Command B referenced the latest request from Command A. The answer from Command C was finally sent to Command B in message ⑦ . However, the additional necessary information requested from Command D in message ② was not received and was finally obtained through a

telephone conversation between Commands B and D. Once all of the necessary information was obtained by Command B, he telephoned the response to Command A and then sent an official confirmation of the call in message (8). This message was sent nearly 19 hours after the previous message and 28 hours after the initial request from Command A. The directive from Command A followed (based upon the information received from Command B) and the remaining messages represent the fulfillment of this directive by subordinate commands.

Figure 2 represents the set of messages for this event which were found at headquarters of Command B. Thus, it is noted that Command B is either an originator or addressee of every message of the set except messages (11) and (14). Message (11) did not appear in the files of Command B (and is so indicated by putting the message number in brackets) although Command B knew of its existence because it was referenced in a later message (number (13)) which he received. Message (14) is another message which was not originally addressed to Command B but which was later readdressed to Command B (represented by dotted lines) by Command C.

Notes on figure 2. By scanning across any row in figure 2, it is possible to see which of the elements of the information available to Command B was also available to the Command represented by that row, as well as the manner in which that Command fits into the general picture. Thus, we notice that Command D is the only one who processed every message found at source B. Note, however, that Command D did not originate any of these messages. Note, also, that Command F entered the picture only as his role became defined. Such interdependence of various command headquarters becomes more evident through a display such as that of figure 2, and the complexity of the communication structure is revealed.[5]

Figure 2 also demonstrates that a message is oftentimes completely meaningless unless the messages which it references are located. This natural application of reference-connected sets to information retrieval may be their greatest potential. It is possible to automatically file messages into appropriate message sets by noting only the references which are given. These sets will then represent events during a crisis and will be available for answering queries regarding the status of events during the crisis. Predetermined subject categories are not required, nor are any restrictions placed upon the format of messages. Thus, the method simply provides a way of quickly locating a message which may have the information (as it is currently expressed in natural language by military commanders) which is necessary to make a decision.

Automatic classification and retrieval. A very simple technique was used in the Dominican Republic analysis for automatically classifying messages into reference-connected sets. It consists of simply filing the events. If a message references a previous message, it is put into the file of the previous message.

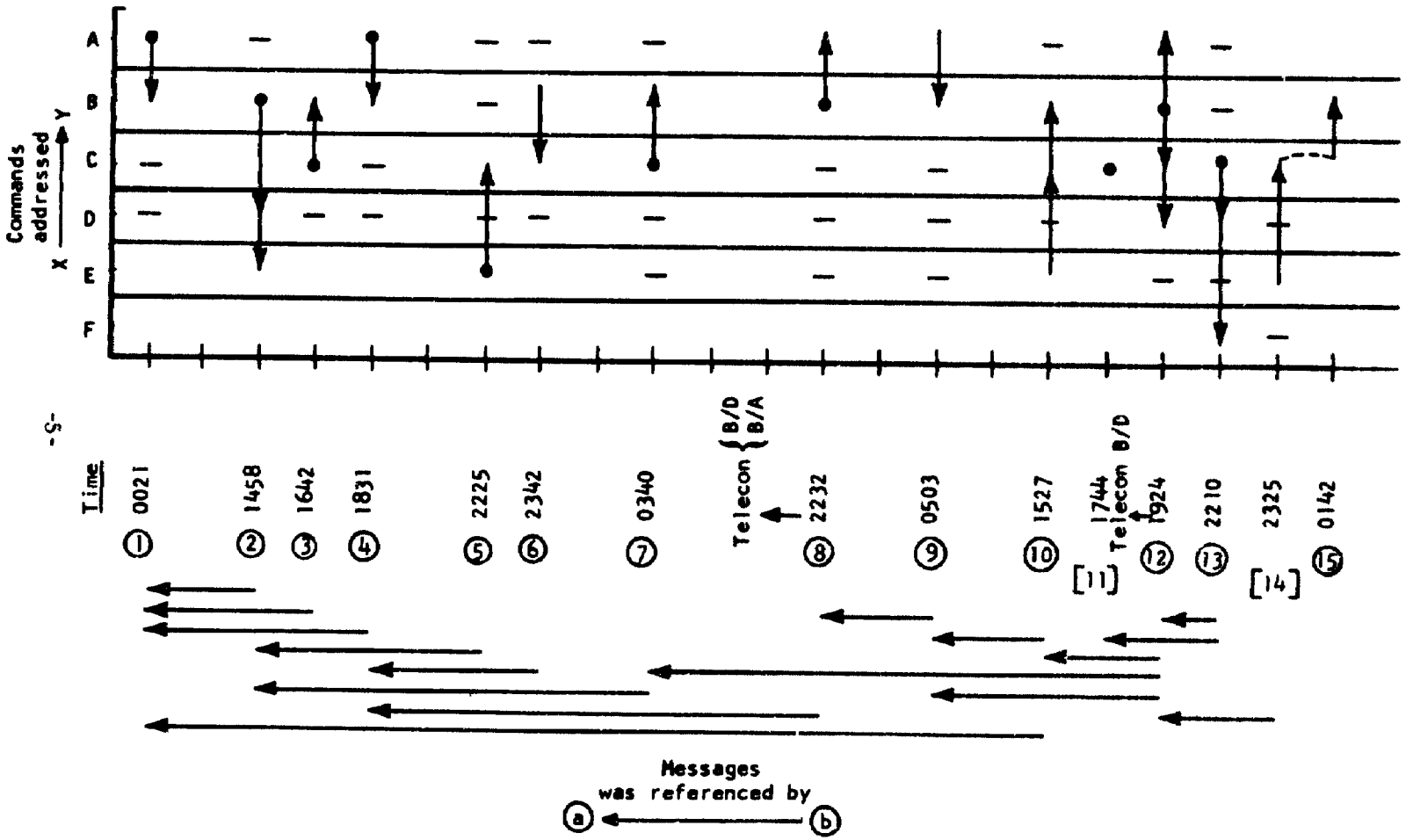


FIG. 2: A REFERENCE-CONNECTED SET

So, for example, in figure 3, message ② would be filed with message ① because it references message ① . Message ③ does not reference a previous message and would thus begin a new file number 2. However, message ④ references messages in both files and therefore connects the two. Two subsets are identified in this way. One subset contains messages ① , ② , and ④ (assigned the number 1). The other subset contains messages ③ , ④ , and ⑤ (assigned the number 2).[6] Message ④ is the link between them and, in the language of directed graph theory, may be considered to be a linking point between two maximal paths in the semipath from message ① to message ⑤ .

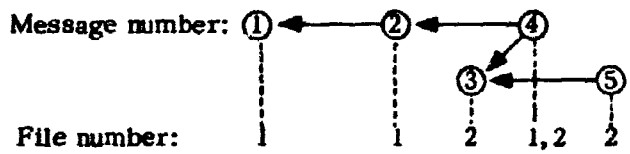


FIG. 3: AUTOMATIC CLASSIFICATION OF A REFERENCE-CONNECTED SET

Subsets. The structure of a reference-connected set identifies subsets (as in the preceding discussion) which can be interpreted in a number of ways. First of all, it is noted that a subset will occur only if there is a message within the set (such as ③ in figure 3) which does not reference a previous received message but which is eventually linked to the set. Such a message may actually begin a "new" event[7] which eventually becomes related in some way to the earlier event initiated by message ① . However, the structure of a reference-connected message set is also a function of another important factor - the organizational chain-of-command and the distribution of information throughout this chain. For a message cannot reference a previous message unless its originator is cognizant of the previous message. Consequently, the paths in a reference-connected set of messages (and thus the corresponding subsets) will often reflect the information flow between specific commands although the event is essentially the same.[8]

For example, in figure 4, message ① is a directive from Command A to Command B which is amplified (and thus referenced) in message ② .

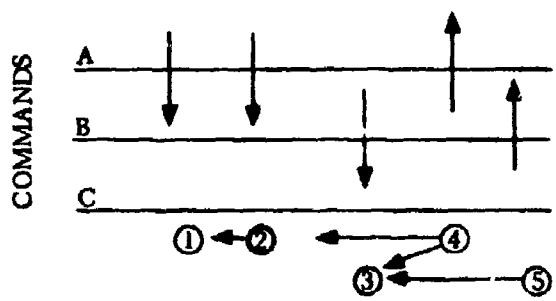


FIGURE 4

Message ③ is a message from Command B to Command C which is based upon the directive in messages ① and ② but which does not reference them. However, sometime later, in message ④, Command B notifies Command A that the directive has been carried out and references message ③. Message ⑤ is the reply from Command C to Command B. In the example of figure 4, note by scanning the rows that each of the three commands has processed a different set of messages. Command A has processed messages ①, ②, and ④ and is aware of message ③ because it is referenced in message ④. Command B has processed all 5 messages. Command C has processed only messages ③ and ⑤. Note also that no subsets occurred in message sets for this "event" at Commands A and C, and that the two subsets at Command B correspond to the sets of messages found at Commands A and C.

It is interesting to note that if Command B had referenced the directive when passing it on to Command C (see figure 5), there would have been no subsets for Command B and Command C would have been aware of an additional message.[9] Finally, if Command B addressed Command A without referencing his communications with Command C (see figure 6), the set would have split entirely for Command B.

FIGURE 5

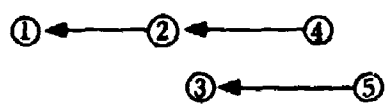


FIGURE 6

It is easily seen that this discussion becomes even more complex when additional commands are included, when multiple addressees are allowed, when the organizational chain is modified, and when the time factor (of which the structure of a reference-connected set is a major function) is considered. Nevertheless, the above discussion makes two points: 1) the subsets of a reference-connected set often corresponded operationally to the sets of messages received at various nodes for the same event (a more thorough study of such subsets may define "domains of information" represented by message traffic at various command headquarters[10]); 2) it is conceivable that rules of referencing might be established which should guarantee that certain kinds of sets and subsets will appear (see appendix A). The possibilities of application of the theory are enormous and this paper will not attempt to consider all of the ramifications.

General application. If the concept of reference-connecting is to be applied to a general information classification and retrieval system, such as a library system, the structural properties become extremely critical. This is because it is not expected that there would be well defined sets of documents dealing with certain subjects as there are of messages dealing with events during a crisis. In the former case, a "new" event or subevent is begun when an initiating message is sent which does not reference a previous message. But documents do not generally behave in this way and, in fact, tend to reference more excessively. However, it is just this concentration upon the evolution of a subject, rather than upon a description of the subject, that is so appealing.[11] Numerous researchers in the field of information retrieval and documentation have been attempting to automatically index, classify, and abstract documents, and a large number of devices have been developed which mechanically or electronically select information from a store. But these devices have been, on the whole, unsuccessful, since the problem of describing the subject covered in a document is even more elusive to machines than it is to humans. Even if a satisfactory classification scheme could be found, there is still the problem of integrating new subject matter into a predetermined code. The reference-connecting technique avoids both of these problems by tracing the evolution of the subject as represented by referenced documents upon which any "new" information is based. It is thus a dynamic, user-oriented system which would have the additional advantage of reflecting the information flow among scientists in any field.[12]

The concept of a dynamic, user-oriented information system is not new, nor is the idea of relating documents through their citations. In 1945, Dr. Vannevar Bush suggested that an individual's personal information storage and selection system could be based on direct connections between documents instead of the usual connections between index terms and documents. These direct connections were to be stored in the form of trails through the literature. Then at any future time the individual himself or one of his friends could retrieve this trail from document to document without the necessity of describing each document with a set of descriptors or tracing it down through a classification tree (reference (d)). In 1956, Dr. R. M. Fano suggested that a similar approach might prove useful to a general library and proposed that documents be grouped on the basis of use rather than content (reference (e)). Dr. M.M. Kessler suggested a criterion for such grouping of technical and scientific papers through "bibliographic coupling"(see reference (c) and footnote 3). A recent research endeavor into search procedures based upon this measure of relatedness between documents used a partitioning model in which sets of highly inter-related documents were found with about 64 - 90 percent retrieval efficiency.[13] The paper states in its concluding section that the model of "trails of documents" as suggested by Dr. Bush has useful features which the partitioning model does not offer. The author says, "Actually, both models have useful features. In some cases there is a definite pattern or trail which should be followed in consulting the documents related to a given subject. In

other cases the order in which the documents should be examined is apparent from their publication data. In still other cases, there is no particular order in which the documents need be consulted. Thus, it would seem that one might want to include both the ideas of sets of documents and trails of documents in a more general information retrieval model (reference (f))."

The tools of directed graph theory applied to the concept of a reference-connected set offers a tempting approach to the information retrieval problem. It still remains to be tested and evaluated for general information systems.

References: (a) Second Congress of the Information System Sciences, editors Spiegel and Walker, 1965

(b) Kessler, M.M., "The MIT Technical Information Project," Physics Today, vol. 18, no. 3, pp. 28-36, Mar 1965

(c) Kessler, M.M., "Bibliographic Coupling between Scientific Papers," American Documentation, vol. 14, no. 1, pp. 10-25, Jan 1963

(d) Bush, Vannevar, "As We May Think," The Atlantic Monthly, vol. 176, no. 1, pp. 101-108, Jul 1945

(e) Fano, R.M., Documentation in Action, chapter XIV-e, pp. 238-244, Reinhold Publishing Corp., New York 1956

(f) Ivie, E.L., "Search Procedures Based on Measures of Relatedness between Documents," Project MAC, MIT, Jun 1966

(g) Harary, F., Norman, R., and Cartwright, D., Structural Models, John Wiley & Sons, Inc., New York, 1965

FOOTNOTES

1. Ruth M. Davis, in a paper presented to the Second Congress of the Information System Sciences (reference (a)) questions whether there is indeed such a science. She notes "... in the volume resulting from the First Congress of Information System Sciences only fourteen technical-type references were cited and three of the nine papers in the volume contained no such references. Scientific doctrine is not built in this manner. As these scientifically accepted procedures become common it will be possible to decide whether information system science will really take its place as a recognized science following respected traditions or whether the introduction of 'information systems' as a formal entity into organizational structures is just a new scientific achievement or a new technological advance."

2. Note that the two concepts - classification and retrieval - are closely interconnected here and elsewhere in this paper. Julian Bigelow states in reference (a) that " inability to carry out the indeterminant activity of abstracting meaning from common language is --- an almost crippling circumstance. As a result, it is not only impossible in general to compare that which may be expressed in search questions to that which may reside in stored material within high speed computational hardware so as to derive a measure of 'nearness' and to issue brief answers at a bit rate acceptable to humans at the terminae, but as a corollary it is impossible also to 'process' automatically so as to reduce the bulk of stored material in any other way that universally preserves textual meaning relevant to arbitrary enquiries, down to a level compatible with rates of human sensory acceptance." He goes on to say that the design of retrieval systems has veered in the direction of attempted compromise in which an intermediate representation (usually "keywords") has been substituted for the document itself.

3. Figure 1 demonstrates the three basic types of reference-connectivity, all of which have been recognized and used by scientists in the Technical Information Project at MIT. (See reference (b) for a description of the project.) M.M. Kessler, in a pioneering article (reference (c)), which was unknown to this author during the development of this technique), suggested a relationship between scientific papers wherein two papers which cited one or more of the same papers were said to be bibliographically coupled. Studies have been conducted to analyze the characteristics of such bibliographic coupling and have indicated that it constitutes a very "meaningful and important type of relationship between papers." Thus, in figure 1, messages ④ and ⑤ are bibliographically coupled. Another type of coupling occurs if two papers are cited by one or more of the same papers, (e.g., ② and ③) and finally, there is the simple citation relationship between ① and ② , ② and ④ , ③ and ④ , and ③ and ⑤ . The MIT Project, however, uses these three basic types of reference

connectivity as separate partitioning criteria for a retrieval system and does not combine them into a single dynamic system for both classification and retrieval as this author proposes to do.

4. This event is chosen to avoid problems of military security. All messages represented in figure 2 are unclassified. As a further precaution, the commands involved will remain anonymous.

5. Figure 2 suggests that if Command B had requested information from Command C in message ② rather than from Command E (who did not have the information, although he may have been the likely person to have it), or if Command B had included Command C as an addressee in message ② , then messages ③ , ⑤ and ⑥ might have been eliminated. The long 28-hour delay from Command A's initial request for information to the response from Command B was also caused by the fact that Command B depended upon an answer from Command D which never came and which was finally obtained through a telephone call. (It is interesting to note that direct communication between Commands B and D was established just prior to the call.) Command D, on the other hand, was closer to the scene and the mere pressure of events may have contributed to his omission.

6. A unique identifier can be assigned to each message, if desired, which will indicate its location in the subset. For example, message ① might be assigned the number 1.0, message ② , the number 1.1, etc. Key messages can also be found by noting those messages which are referenced most frequently.

7. The definition of an "event" is not given and must be considered to be intuitive in the case of military operations during the Dominican Republic crisis - e.g., sending food, landing troops, providing medical supplies, evacuating personnel, etc.

8. This is again a question of definition. Communications between specific commands in partial fulfillment of a more general directive might also be considered to be a subevent or even a separate event. This becomes a more philosophical problem which will not be dealt with here and which is avoided in the more general discussion below by emphasizing paths from document to document rather than the partitions between them.

9. This was the case in the second message of the event illustrated in figure 2. Command B referenced the first message from Command A in passing the request for information to Commands D and E. However, in this case, Command A was also addressed in this message.

10. The notion can be pursued further: how much overlapping of information exists at the various headquarters and might this not give a certain measure of the centralization/decentralization of the communication structure?

11. Ruth M. Davis hints that citation characteristics are an indication of how scientific doctrine is "built." See footnote 1.

12. The discussion of the command organizational effect upon reference-connected sets is not entirely inapplicable to the general study. It is still true that a document cannot reference a previous document unless its author is cognizant of the previous document and thus the paths of a reference-connected set will tend to reflect the information flow between certain authors, certain organizations, certain libraries, etc.

13. The most widely accepted method of evaluating the performance of information retrieval systems is currently through the recall and relevance ratios. The recall ratio is the percentage of relevant items that are actually retrieved and the relevance ratio is the percentage of retrieved items that are relevant. In determining what is or is not relevant, recourse is usually made to an indexer or a user. Recent studies have shown that these people are able to agree among themselves as to how documents should be classified in at most 80 percent of the cases. This "failure" of humans to index consistently has led some to try to find better automatic "non-judgemental" standards on which to validate relevance. (Quoted from reference (f).) In the case of message traffic during the Dominican Republic crisis, reference-connected sets included over 90 percent of the messages which pertained to an event. The relevance ratio was simply avoided by defining an event appropriately. Sometimes what intuitively appeared to be two or more events became connected, but they were later found to be related and could still be identified through subsets of the reference-connected set.

(REVERSE BLANK)

APPENDIX A

A COMMUNICATION MODEL FOR MILITARY COMMAND STRUCTURES

Consider a command structure in which there are three commands A, B, and C representing three hierarchical echelons. (Command A is superior to Command B who is superior to Command C.) There are four possible ways in which Command A can direct an action to be performed by Command C through Command B. These four ways are illustrated in figures A-1 through A-4.

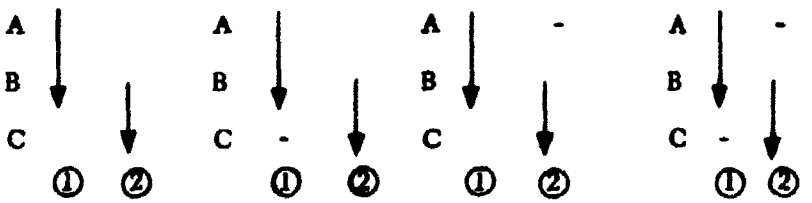


FIGURE A-1 FIGURE A-2 FIGURE A-3 FIGURE A-4

The commands are represented on a vertical axis, and the messages (assumed to be TTY communications although the discussion holds for any flow of information) are numbered on a horizontal axis. A vertical arrow from, say, Command A to Command B indicates that A addressed (↓) B for action. A bar (–) at Command C indicates that C was an information addressee on the message.

In the first case (figure A-1), Command A sends a directive to Command B in message ① . Command B receives the directive, interprets it in light of his responsibilities, makes a decision, and sends a corresponding (usually more specific) directive to Command C in message ② . In the second case (figure A-2), Command A also addresses Command C for information in message ① so that Command C will have previous knowledge of the subject before receiving the directive from Command B. (The case in which Command A addresses Command C for action is not considered, for this would reduce the structure to a simple superior/subordinate relationship between Commands A and C which would require no decision from the intermediate Command B. This sometimes happened during the DomRep crisis in which either the time constraint or political considerations appeared to necessitate such direct communication. Another case arises if the nature of the directive to Command B is such that he can simply readdress or relay the message to Command C. However, in this discussion it is assumed that a strict functional chain-of-command exists in which each member has his own theater of responsibility and makes decisions as partial fulfillment of a general directive.)

In the first two cases, Command A is unaware of the directive which is given in message ② to Command C. Thus, two other cases arise if Command A is an information addressee in message ② . These last two cases (shown in figures A-3 and A-4) have the advantage of requiring no additional message from Command B to Command A which relates that the directive has been carried out.

All of the above cases were evidenced in traffic during the DomRep crisis. As suggested, there are advantages and disadvantages of each, depending upon the criteria used. If reduced message volume is the only goal, the structure of figure A-1 is preferable. (By scanning the three rows representing traffic originated and received by each of the three commands, we see that in figure A-1 Commands A and C process one message each and Command B processes two messages. In figure A-2, Command C processes an additional message, etc.) However, it was already noted that the structure of figure A-1 requires an additional message if Command A is to be kept informed of the directive to Command C. This suggests that structure 3 is preferable, for it informs A while not burdening C (who is a lower echelon command and is often less equipped for high message volumes) with an additional message. On the other hand, structure 3 has the possible disadvantage of not notifying Command C in advance. Another consideration is the fact that in this structure Command C does not know the basis for the directive from Command B. Also, it may happen that in future communications for this directive, the original message from A is referenced in a message addressed to C. It is in this case that rules of referencing may be particularly helpful. If message ② referenced message ① , Command C would have a record of the message without actually receiving it. Furthermore, if message traffic were being automatically filed into reference-connected sets at Command C headquarters, any future messages which reference the original directive from A will then connect to the set. Observe also that Command B must reference message ① in his directive to Command C if he wants to automatically construct a reference-connected message set dealing with this event which will contain communications with both superior and subordinate commands. (During DomRep some commanders did and some didn't reference in this way.) If Command B chooses to separate his communications with higher and lower commands and not reference message ① in message ② , then these two messages will also appear in separate sets for Command A and/or Command C (except in figure A-1, where the communications are the most decentralized in the sense that only the intermediate command processes both messages).

Automatic distribution

There are many other considerations and the logic of distribution becomes even more complex when additional commands and levels are added. Nevertheless, the contribution of reference-connected sets to the establishment of a communication scheme as well as to the possible automation and interface of information flow is evident. There are other possibilities as well. For example, reference-connected sets may provide a solution to the problem of distribution logic. It is conceivable that the distribution of a message could be automatically

determined if reference-connected sets represent events and if previous communications dealing with an event suggest the commands which should be informed. For example, in figure A-3, if Command C is required to send a further directive to another Command D, he will know from message ② that Command A should also be informed. Figure A-5 illustrates such an extension of figure A-3. Horizontal arrows between messages indicate that message ⓧ is referenced by (◄—) message ⓨ .

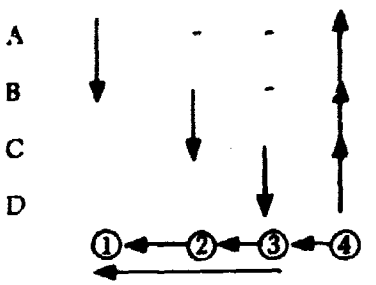


FIGURE A-5

Advantages

In figure A-5, the directives down are "decentralized" in that each command has the responsibility of directing his chosen subordinate. But the report of the final results to higher commands (who have monitored the progress as information addressees) can be "centralized" if the situation requires. All addressees except the new action addressees are automatically determined by the previous message. The volume of traffic processed at each command headquarters decreases as the command level decreases. However, each command has a record (through the formal references) of all previous communications dealing with this event. In fact, all of the messages automatically form a reference-connected set at each command node. This is because a message references not only the previous message, but also the references which that message contains. Thus, even the references are determined automatically. There are no interface problems and the scheme can be easily adapted to other command structures.

Naturally, there will be "non-decision" messages exchanged between commands which are amplifications, clarifications, recommendations, requests for information, advice, etc. Such communications would not follow the scheme of figure A-5 but would rather address only the commands involved and would reference the messages to which they are most directly related. (See figure A-6.)

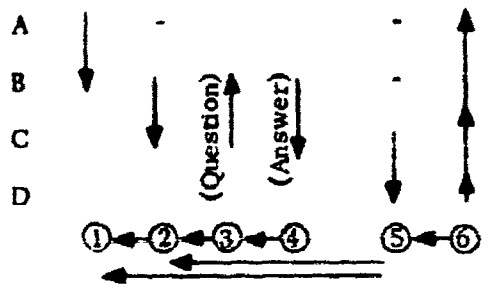


FIGURE A-6

CONCLUSIONS AND RECOMMENDATIONS

The structure of a reference-connected message set is a function of the organizational chain-of-command and reflects the distribution of information throughout this chain. Consequently, reference-connected sets are useful in developing models of command information flow. Furthermore, a simple rule of referencing can be established which should guarantee that all messages will fall into appropriate sets at each command node. Such a rule can also aid in automatic determination of messages addressees and can contribute to the establishment of a communication scheme which is more amenable to automation.

The proposed communication model is theoretical and must be adapted to various command structures in operational situations. It is recommended that its potential be examined through an operational trial.

None

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Operations Evaluation Group, Center for Naval Analyses, of the Franklin Institute	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP None

3. REPORT TITLE

A Reference-Connecting Technique for Automatic Information Classification and Retrieval

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Research Contribution

5. AUTHOR(S) (First name, middle initial, last name)

Greene, M.J.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
10 March 1967	18	7

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-C-0211	Operations Evaluation Group Research Contribution No. 77
b. PROJECT NO -------	
c. -------	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None
d. -------	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	Office of Naval Research Department of the Navy Washington, D.C. 20350

13. ABSTRACT

A recent study of command information flow associated with the Dominican Republic coup of April-May 1965 introduced an analytical tool for identifying deficiencies in the flow and use of information which appears to have considerable potential as a general technique for information retrieval. Naval messages are associated with each other through their formal references in a manner analogous to the concept of "joining" as defined in a recent publication in directed graph theory by Harary, Norman, and Cartwright (Structural Models). "Reference-connected sets" are then constructed from message traffic dealing with the coup and are found to uniquely identify operational events during the crisis. Such sets (and subsets if further refinement is desired) can be obtained in real-time and a very simple method is demonstrated which automatically classifies messages as they enter the system. This technique, if applied to a library system, avoids both the problem of describing the subject covered in a document and the problem of integrating new subject matter into a predetermined classification code. It is a user-oriented system which combines static coupling methods used in the Technical Information Project at MIT and converts them into a dynamic process for both classification and retrieval. Thus, it concentrates on the evolution of the subject as represented by referenced documents upon which any "new" information is based and it has the additional advantage of reflecting the information flow among scientists in any field.

None

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
information classification and retrieval reference-connected sets subsets chain-of-command user-oriented information system command information flow						

DD FORM 1 NOV 65 1473 (BACK)

(PAGE 2)

None

Security Classification